

Remarks

Regarding the requested amendments to Specification

In the Final Office Action the Examiner objected to the Specification (see point 1, page 2 of the Office Action) because of the informality that the information in the first paragraph of the application needed to be updated to indicate that the parent '768 and '068 applications are now abandoned. The presently requested amendment to the Specification cures this defect and informality. In addition, the new paragraph now also correctly indicates that application 09/623, 068 is a National Stage Entry under 35 U.S.C. 371 of, and claims priority from PCT/US99/04376 (see USPTO File Wrapper: Query Control Form dated 3/17/2004 and Correction of Bibliographic Data on page 1 of Specification dated 3/26/2004).

Regarding new paragraph [0026.1] that is requested to be added to the Specification.

The addition of this paragraph is requested to provide more literal antecedent basis in the Specification for a limitation that is already present in pending, previously allowed claims. As noted in MPEP 2173.05(e), it is not necessary that claim terms or phrases have literal antecedent basis in the specification. Nevertheless applicants respectfully request the addition of this paragraph. The paragraph deals with conventional linkage study techniques that are essentially one-dimensional and essentially one-dimensional marker panels. *"An essentially one-dimensional panel of markers for a linkage study"* is any conventional (at the time of filing of the first priority document, US Provisional 60/76, 102) linkage study marker panel chosen to attempt to achieve one-dimensional closeness and linkage of one or more panel markers and the (or a) sought trait-causing polymorphism.

The subject matter of the requested paragraph [0026.1] is not "new matter". The basis for the addition of paragraph [0026.1] to the Specification is in MPEP 2163.06 and in MPEP 2163.07. The first paragraph of MPEP 2163.06 states: *"information contained in.. the specification ..as filed may be added to any other part of the application without introducing new matter"*. And relevant sections of MPEP 2163.07 (Amendments to Application Which Are Supported in the Original Description) are I. REPHRASING and Inherent Function, Theory or Advantage.

BEST AVAILABLE COPY

It is well-known in the art of linkage studies that linkage study "markers are chosen based on a principle of one-dimensional closeness" in the chromosomal location dimension. This chromosomal closeness or nearness is cited in the present application at bottom [0009], top [0011] and bottom [0016]. Specifically, "*By establishing linkage, especially strong linkage, between a known marker and an unknown gene [or trait-causing polymorphism] it is possible to locate the gene [or trait-causing polymorphism] near to the chromosomal location of the known marker.*" And "*Linkage studies are a method of establishing linkage between a marker and a gene [or trait-causing polymorphism] or genes.*" "*Strong positive evidence for linkage of the markers (from the scanned chromosomal region) to a gene or genes responsible for a characteristic or trait is strong evidence that a trait-causing gene or genes is located within the chromosomal region.*" (The application uses the terms "gene" and "trait-causing polymorphism" interchangeably, see [0005] and Definitions Section [0059].)

The concept of chromosomal closeness or nearness is also present in the term "linkage". See enclosed definitions of linkage. Specifically the Genome Glossary of the U.S. Government's Human Genome Project (http://www.ornl.gov/sci/techresources/Human_Genome/glossary/glossary.shtml#L) dated 1/30/2007 defines "*Linkage The proximity of two or more markers (e.g., genes, RFLP markers) on a chromosome; the closer the markers, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in prokaryotes, mitosis or meiosis in eukaryotes), and hence the greater the probability that they will be inherited together.*" See also the enclosed, marked pages of the Encyclopedia of Molecular Biology and Medicine (1996) editor Robert A. Meyers. On page 377, volume 3, Linkage (of genes) is defined as: "*The tendency of genes to be inherited together based on proximity within the same chromosome*". And on page 222 Volume 1 under Linkage Analysis the Encyclopedia says: "*When a gene...is located on the same chromosome pair as marker and close to it (i.e., when gene and marker are linked)..*"

Similarly, for example, the bottom of [0035] of the present application refers to one-dimensional closeness in a one-dimensional view (or perspective). And non-limiting examples of conventional essentially one-dimensional linkage study scanning techniques given in the application (see for example [0020]) use a strategy that attempts to locate at least one marker near the (or a) sought trait-causing polymorphism (in a chromosomal region) by distributing markers approximately evenly (along the length of the chromosomal region). Another non-limiting example of a conventional essentially one-dimensional linkage study, based on one-dimensional closeness, is the example given in mid [0027], specifically the TDT association study of Risch and Merikangas. The TDT association study of Risch and Merikangas is based on the optimal assumption of the analyzed allele being the disease allele (i.e., the assumption of a study marker being the disease-causing polymorphism). (As is well-known, in the case of association studies, markers are chosen to attempt to achieve closeness, linkage and linkage disequilibrium between one or more of the markers and the (or a) trait-causing polymorphism.)

The favoring in conventional linkage study techniques for markers with least common allele frequencies near 0.5 is discussed in [0026]. However, the application repeatedly emphasizes that comparatively little attention is paid to the allele frequency dimension, see for example, [0023], [0024], middle [0026], top [0035]. This comparatively little attention is because (as stated in new paragraph [0026.1] and the Amendment/Response of December 2004) *“Conventional, essentially one-dimensional marker panels are.... not chosen based on using the principle of the similarity of marker allele frequency and possible trait-causing polymorphism allele frequency to increase the power of an association-based linkage test to detect evidence for linkage”*.

This limitation is present in pending, previously allowed claims and was discussed on page 11 of the Amendment/Response of December 2004. For the Examiner's convenience, that discussion is reproduced here and stated the following: see, e.g. [0019], [0020], top [0024] and [0035] (i.e., conventional linkage study techniques are essentially one dimensional, focus on the dimension of chromosomal location but give little attention to the dimension of allele frequency) and see, e.g. [0308] “It is well known that increased disequilibrium between a marker and linked disease locus increases evidence for linkage provided by association-based linkage tests such as the TDT. However, what has not been recognized is that the specific allele frequencies of the marker locus can also have an enormous impact on the strength of evidence for linkage.” And see a rendition of the principle that the inventor discovered: e.g. [0285] i.e., the power of association-based tests for linkage are increased as the allele frequencies of the disease-causing (or trait-causing) allele of a bi-allelic gene (or polymorphism) and a positively associated allele of a linked bi-allelic marker become similar in magnitude. That is, conventional (essentially one-dimensional) techniques are not based on using similarity of marker allele frequency and possible trait-causing polymorphism allele frequency to increase the power of an association-based linkage test to detect evidence for linkage.

As stated in the Amendment/Response of December 2004, the reason for the comparative inattention to allele frequency is because the above principle (of the similarity of marker allele frequency and possible trait-causing polymorphism allele frequency increasing the power of association-based linkage tests) was discovered by the inventor and was unrecognized by conventional techniques; see, for example, [0308] and top [0285]. The principle was, for example, unrecognized at the time of the conventional, essentially one-dimensional TDT association study of Risch and Merikangas in September of 1996, see [0027]. There is nothing mentioned about the principle in the Risch and Merikangas reference or in the Kruglyak reference [0026] (Nature Genetics September 1997). The inventor's original manuscript was first submitted for publication in December of 1996 (see [0285]), but was not published until 1998.

Regarding the amendments to previously pending claims

The limitation *"wherein the group of two or more covering markers is not an essentially one-dimensional panel of markers for a linkage study, wherein the essentially one-dimensional panel is a panel not based on using similarity of marker allele frequency and possible trait-causing polymorphism allele frequency to increase the power of an association-based linkage test to detect evidence for linkage"* has been expressly added to each of the previously pending, allowed independent claims 6, 16, 39, and 44 and to previously pending claim 61. The limitation was already present in previously pending, allowed dependent claims 7, 17, 40, 45 and in previously pending dependent claim 62. Each of currently amended claims 6, 16, 39, and 44 continue to be within the scope of previously allowed claims 6, 16, 39, and 44. The limitation is discussed in detail above in relation to new paragraph [0026.1]. The limitation was also previously discussed in the Amendment/Response of December 2004, p. 11, see above.

Claim 15 has been cancelled and a similar new claim, new independent claim 81 has been submitted. Claim 15 was not in exact conformity with the description at [0169] and [0170]. New claim 81 is in closer conformity with the description at [0169] and [0170]. The limitation *"wherein the localizing uses a technique or techniques that detects gradients, wherein the detection technique or techniques uses a gradient along the allele frequency dimension"* was present in previously allowed claim 15. The limitation was discussed in the Remarks on page 13 of the December 2004 Amendment/Response. Those Remarks state: *"Regarding support see, e.g. [0169], [0170], [0171]. See also, e.g. [0285] through [0289] inclusive and [0296] which describe increases in power along the allele frequency dimension, i.e. one or more gradients in power along the allele frequency dimension."* (Power and statistical evidence for linkage are closely related or equivalent, see [0286].)

The invention of claim 15 is a two-dimensional linkage study technique and is based (as are all the other claimed inventions) on the inventor's newly discovered principle of the similarity of marker allele frequency and possible trait-causing polymorphism allele frequency increasing the power of association-based linkage tests. Conventional linkage studies do not make use of a gradient in statistical evidence for linkage or power along the allele frequency dimension. Applicants respectfully submit that the invention is novel and unobvious by virtue of the limitation *"wherein the localizing uses a technique or techniques that detects gradients, wherein the detection technique or techniques uses a gradient along the allele frequency dimension"*.

New claim 82 depends from new claim 81. Like most of the other claimed inventions, the marker panel used in the invention of claim 82 is expressly defined as "*not an essentially one-dimensional panel for a linkage study*". Put another way, new claim 82 comprises steps a), b), c), d) and e) of the claimed process of claim 6. So claim 82 is equivalent to: ***A process for identifying one or more bi-allelic markers linked to a bi-allelic trait-causing polymorphism in a species of creatures, comprising acts of: a), b), c) d) and e) of claim 6, further comprising the act of: f) localizing the trait-causing polymorphism to the chromosomal location-least common allele frequency (CL-F) location of one or more markers that show evidence for linkage based on the calculations of act e), wherein the localizing uses a technique or techniques that detects gradients, wherein the detection technique or techniques uses a gradient along the allele frequency dimension.***

Claim 57 has been amended and the identifier "a)" has been eliminated from the claim. Applicants respectfully submit that the amendment is a mere informality and does not change the scope of the claim. Applicants believe that the amendment increases claim clarity as the identifiers "a)" and "b)" are used in independent claim 39 from which claim 40 and claim 57 depend.

Claims 61-67 were rejected in the Final Office Action of Oct 2006 as indefinite under 35 USC 112, 2nd paragraph. The rejection referred to MPEP 2172.01 and the omission of critical steps. An Examiner Interview regarding the rejection of these claims was conducted on December 21, 2006 and applicants respectfully offered the following arguments with respect to patentability of these claims. Applicants argued that MPEP 2172.01 also cites MPEP 2164.08(c), which states "*Features which are merely preferred are not to be considered critical*". And applicants presented support in the description for practicing the claimed processes without the minimal steps recited in claim 39.

More specifically, applicants cited parts of the description which indicate the processes can be practiced by a computer without the minimal steps recited in claim 39. (It is possible, for example, to obtain genotype data/sample allele frequency data from a stored computer file without working directly with chromosomal DNA.) Paragraph [0173] states: "*It is also possible for a computer program to execute any one of the steps or step-like parts of Process#1*". And paragraphs [0208] and [0209] recite an appropriately programmed computer as an example of a means to obtain genotype data/sample allele frequency data of step d) of Process#1. And paragraph [0231] refers to a process to obtain genotype data/sample allele frequency data similar to the data of d) of Process #1.

The claimed processes in claims 61-67 are essentially processes for practicing the step-like part of obtaining genotype data/sample allele frequency data recited in d) of Process#1, see [0155] and [0152] and [0154]. Applicants respectfully submit that these processes are novel and unobvious because the group of two or more markers which systematically cover a CL-F region (which are also not an essentially one-dimensional panel) are essentially novel and unobvious.

Applicants respectfully submit that the terms "genotype data" and "sample allele frequency data" are definite and are well-known in the art. In addition, the related term "genotype data/sample allele frequency data" is specifically defined in the application (see [0148] and related paragraph [0147]).

Applicants respectfully submit that the act of "obtaining" such data for use in a linkage study is definite (see [0166]).

McMahon, et. al. Integrating Clinical and Laboratory Data in Genetic Studies of Complex Phenotypes; A Network-Based Data Management System (American Journal of Medical Genetics, 81: 248-256 (1998)) is a reference that was published about the time of filing of the parent PCT application and Provisional priority applications. The reference describes a computer-based system for storing, managing and accessing genetic data, including genotype data. Some marked pages from the reference are submitted to the Examiner as illustrative of knowledge in the art of using previously collected genotype data in linkage studies.

The inventor's paper in the Annals of Human Genetics (see [0029] and footnote 11) is an integral part of the application and is incorporated by reference into the application (see [0333]). The first page of this paper refers to three linkage studies (Julier et. al., Spielman, et. al., and Thomson, et. al.) that used previously collected genetic data, including genotype data, from Genetics Analysis Workshop 5 (GAW5) families. Some marked pages from the inventor's paper in the Annals of Human Genetics (AHG98) and from the Julier, Spielman, and Thomson references are submitted to the Examiner as illustrative of knowledge in the art of using previously collected genotype data in linkage studies.

Some further facts about the Examiner Interview As related to the Examiner in the December 21, 2006 interview, the Kruglyak reference (Nature Genetics September 1997 [0026]) is an example of a conventional, essentially one-dimensional linkage study technique or approach. As related to the Examiner in the December 2006 interview, the inventor's original unpublished manuscript (first submitted for publication in December of 1996, see [0285]) was, at first, rejected for publication. The unpublished manuscript contained concepts about the unrecognized importance of allele frequency (see for example mid to bottom [0031], [0290], [0293], [0300]).

New, dependent claims 83-90 have been submitted These new claims are similar to previously allowed claims and all have limitations present in previously allowed claims. More specifically new claims 83, 86 and 89 are similar to allowed, pending claims 17, 46 and 69 respectively. But these new claims do not have the limitation, “*wherein each bi-allelic covering marker is an exact, true bi-allelic marker*”. As previously discussed on p. 13 of the Amendment/Response of May 30, 2006, the term “bi-allelic markers” in the art generally means exact, true bi-allelic markers. (For example, SNPs are examples of such exact, true bi-allelic markers.) The specification, however, also expands the term “bi-allelic marker” somewhat and describes bi-allelic marker equivalents or BMEs (mathematical markers formed from one or more markers that act like they are bi-allelic) and approximate bi-allelic markers, see for example [0054] and [0055].

Correction of Attorney misstatements in the record The applicants now correct two misstatements in the Remarks of previously filed responses made by the applicants' attorney. These corrections are made to avoid any future confusion. (As the Court in Biotec Biologische vs. Biocorp (249 F 3d 1341, 58 USPQ2d 1737) essentially noted, attorney errors in the prosecution record must be viewed in context and be considered in light of other statements in the same prosecution record.)

Misstatement 1: The applicants' attorney has previously made the following misstatement in the Remarks Section of previously filed responses (p. 11 December 2004 & p. 20 August 2006): “*For the record, the applicants note that the linkage disequilibrium in the well known principle quoted above from [0308] is essentially measured in a specific way: i.e. the increased disequilibrium is computed respectively as δ/δ_{\max} for $\delta \geq 0$ or δ/δ_{\min} for $\delta < 0$, wherein each of the δ values is a value of the coefficient of disequilibrium. This is the way (or essentially the way) that increased linkage disequilibrium is computed in the application in paragraphs [0291], [0292], in Table 2 on page 21, in AHG 98 in Tables 1, 2, and 3 pp. 165, 167.*” This, however, is a misstatement.

Clarification Paragraph [308] of the application states: “*It is well known that increased disequilibrium between a marker and linked disease locus increases evidence for linkage provided by association-based linkage tests such as the TDT.*” The applicants now make the following clarification. It is true that such a concept as stated above in [308] was well known in the art. And this concept is consistent with the findings in the inventor's paper (Annals of Human Genetics, 1998, vol 62, pp. 159-179, referred to herein as AHG98). These findings in the inventor's paper (AHG98) include (p. 160): “*(2) TDT power is increased by disequilibrium between a bi-allelic marker and disease locus*”.

It is also true that in AHG98, disequilibrium, including increased disequilibrium, is measured (or essentially measured) in a specific way: i.e. the disequilibrium is computed respectively as δ/δ_{\max} for $\delta \geq 0$ or δ/δ_{\min} for $\delta < 0$, wherein each of the δ values is a value of the coefficient of disequilibrium. However, the statement of the well-known concept in [308] of the application does not necessarily mean that the disequilibrium (or increased disequilibrium) in the concept is measured (or essentially measured) as δ/δ_{\max} for $\delta \geq 0$ or δ/δ_{\min} for $\delta < 0$. Such a statement is not made in the originally filed application. Applicants' attorney's statements to this effect in the earlier cited Remarks (p. 11 December 2004 & p. 20 August 2006) were erroneous. Moreover, the findings in AHG98 are not admitted to being prior art with respect to the present invention by any Remarks made or by paragraph [308].

Misstatement 2: The applicants' attorney has previously made the following misstatement in the Remarks Section of previously filed responses (bottom p. 20 August 2006): "***The linkage disequilibrium (including increased linkage disequilibrium) which essentially one-dimensional panels attempt to achieve is measured (or essentially measured) as δ/δ_{\max} when $\delta \geq 0$ and δ/δ_{\min} when $\delta < 0$.***" Again, this statement is erroneous. As stated above the inventor's findings in AHG98 are consistent with the well-known concept in [308], but the linkage disequilibrium (including increased linkage disequilibrium) which essentially one-dimensional panels attempt to achieve is not necessarily measured (or essentially measured) as δ/δ_{\max} when $\delta \geq 0$ and δ/δ_{\min} when $\delta < 0$. And the term "*an essentially one-dimensional panel of markers for a linkage study*" should not necessarily be construed to include the measurement limitations above involving δ/δ_{\max} when $\delta \geq 0$ and δ/δ_{\min} when $\delta < 0$. Such a statement is not made in the originally filed application. Rather, "*an essentially one-dimensional panel of markers for a linkage study*" is any conventional (at the time of filing of the first priority document, US Provisional 60/76, 102) linkage study marker panel chosen to attempt to achieve one-dimensional closeness and linkage of one or more panel markers and the (or a) sought trait-causing polymorphism. Such essentially one-dimensional panels are described above, including in new paragraph [0026.1].

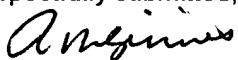
Conclusion

All points of rejection and objection in the Final Office Action of October 3, 2006 have been answered. In this RCE & Amendment/Response the applicants have submitted an amended first paragraph of the Specification and have requested the addition of new Specification paragraph, [0026.1]. Some claims have been amended as discussed above, 1 claim has been cancelled, 1 new independent claim and 9 new dependent claims have been added.

Appropriate small entity fees for a 1 month extension, an RCE, and extra claim fees for 1 new independent and 9 new dependent claims are also enclosed.

For the reasons advanced above, applicants respectfully submit that the application is now in condition for allowance and that action is earnestly solicited.

Respectfully submitted,



Robert O. McGinnis

Registration No. 44, 232

February 5, 2007

1575 West Kagy Blvd.

Bozeman, MT. 59715

tel (406)-522-9355

Enclosures: Selected pages from the McMahon, the Inventor's paper AHG98, Julier, Spielman, and Thomson references, some marked. Total of 15 pages.

Integrating Clinical and Laboratory Data in Genetic Studies of Complex Phenotypes: A Network-Based Data Management System

Francis J. McMahon,^{1*} C.J.M. Thomas,¹ Rebecca J. Koskela,² Theresa S. Breschel,¹ Tyler C. Hightower,¹ Nichole Rohrer,¹ Christine Savino,¹ Melvin G. McInnis,¹ Sylvia G. Simpson,¹ and J. Raymond DePaulo¹

¹Department of Psychiatry and Behavioral Sciences, The Johns Hopkins University School of Medicine, Baltimore, Maryland

²Department of Genetics, Stanford University, Stanford, California

The identification of genes underlying a complex phenotype can be a massive undertaking, and may require a much larger sample size than thought previously. The integration of such large volumes of clinical and laboratory data has become a major challenge. In this paper we describe a network-based data management system designed to address this challenge. Our system offers several advantages. Since the system uses commercial software, it obviates the acquisition, installation, and debugging of privately-available software, and is fully compatible with Windows and other commercial software. The system uses relational database architecture, which offers exceptional flexibility, facilitates complex data queries, and expedites extensive data quality control. The system is particularly designed to integrate clinical and laboratory data efficiently, producing summary reports, pedigrees, and exported files containing both phenotype and genotype data in a virtually unlimited range of formats. We describe a comprehensive system that manages clinical, DNA, cell line, and genotype data, but since the system is modular, researchers can set up only those elements which they need immediately, expanding

later as needed. *Am. J. Med. Genet. (Neuropsychiatr. Genet.)* 81:248-256, 1998.

© 1998 Wiley-Liss, Inc.

KEY WORDS: relational database; linkage; data integrity; modular

INTRODUCTION

The identification of genes underlying a complex phenotype can be a massive undertaking. Data management for such studies must cope with large sample sizes, multiple data storage sites, and some data that change over time. The integration of such large volumes of clinical and laboratory data has become a major challenge in genetic studies of complex phenotypes.

Gene identification in complex phenotypes may require a much larger sample size than thought previously. Large sample sizes may be needed for the initial detection of linkage, and even larger sample sizes for replication of linkage findings [Suarez et al., 1995]. Narrowing the linkage finding to a physically-mappable chromosomal location may require more than 2,000 affected sib-pairs [Kruglyak and Lander, 1995]. Furthermore, each affected subject is typically associated with a large amount of clinical data and the more than 300 genotypes that are generated in genome-wide linkage searches.

Each type of primary data has special storage requirements. Clinical assessments are typically collected on handwritten forms and are often supported by copies of medical records and other documentary evidence in various nonstandard formats. Blood samples and cell lines must be tracked from subject to freezer. Genotype data may exist in the form of autoradiographs or the specialized data files produced by automated genotyping systems.

The data generated in genetic studies are not static, but change over time. Previously unavailable relatives may volunteer for the study or previously studied subjects may die. Clinical data may change after longitu-

Contract grant sponsor: Charles A. Dana Foundation; Contract grant sponsor: NIMH; Contract grant sponsor: National Alliance for Research on Schizophrenia and Depression; Contract grant sponsor: Johns Hopkins University Affective Disorders Fund; Contract grant sponsor: Johns Hopkins University George Browne Laboratory Fund.

*Correspondence to: Francis J. McMahon, M.D., Department of Psychiatry and Behavioral Sciences, The Johns Hopkins University School of Medicine, Meyer 3-181, 600 N. Wolfe Street, Baltimore, MD 21287-7381. E-mail: fmcm@welchlink.welch.jhu.edu

Received 10 September 1997; Revised 13 January 1998

dinal follow-up. DNA sample supplies dwindle and must be replaced. Genotype data may require correction if they fail to segregate or when false paternity or sample mix-ups are detected. Thus, regular archiving and updating of data are required to forestall degeneration of the database over time.

In this paper we describe a network-based data management system designed to address the special requirements of family studies of complex phenotypes. The system is expandable, modular, and easily adapted to a wide variety of studies. The system uses existing relational database, pedigree, and networking software and standard PC hardware. The efficient integration of clinical and laboratory data in the form of output files, summary reports, and pedigrees is a major feature.

MATERIALS AND METHODS

User Input

Our first task in designing a data management system focused on the "users," i.e., the clinicians, research assistants, and laboratory technicians who would be using the system every day. We involved the users in deciding which data elements needed to be considered, the naming of fields and tables, and the design of data entry forms. After each module of the system was made available, users were polled in a series of feedback meetings about whether that module was accomplishing the desired tasks in an efficient and user-friendly fashion. If not, that module was redesigned to better fulfill user needs. After the system was entirely in place, further expansions and modifications were considered as needed.

Hardware Requirements

The hardware requirements for this system depend on the size of the sample to be studied and the modules used. At least one computer is required to house the main database, and (ideally) at least one computer is devoted to each module, located in a spot where the users will have ready and continuous access. As the main computer we use a Pentium 75-MHz machine (Compaq Computer Corp., Houston, TX) with 16 MB of RAM, a 2-gigabyte hard drive, and a Colorado Jumbo 1400 tape drive (Hewlett Packard, Palo Alto, CA). These should be viewed as reasonable minimum requirements, since this computer stores all the data and acts as a server for the entire system. For each module we use at least one 486/66-MHz or Pentium 75-MHz computer with at least 16 MB RAM and a 500-MB hard drive. For the genotype module, a Macintosh computer (Apple Computer, Cupertino, CA) is also needed if automated genotypes will be processed using the GeneScan 2.1fc2 and Genotyper 1.1 programs (Applied Biosystems, Foster City, CA).

If additional computers are being used, then a network must also be in place and each networked computer needs a network card. The network system we use is a departmental local area network (LAN), connecting each computer to a central hub, with structured cabling utilizing 10BaseT Ethernet. The hub also connects the LAN to the campus-wide network using fiber-optic cable, which provides access to the Internet.

Each computer should be connected to a printer, either through the network or through a printer buffer. We use two different types of printers. Our main printer is a Hewlett Packard (HP) LaserJet 4M Plus (Hewlett Packard, Palo Alto, CA), which is linked to the network and is centrally located for multiple users. We also have HP Inkjet 500-series printers attached directly to the computers that serve the DNA, cell line, and genotype modules.

The DNA module is designed to work with a spectrophotometer that measures DNA concentration and quality. We use an HP Diode Array Spectrophotometer (model 8452A) connected to an HP Vectra 486/33N computer with 16 MB RAM. Included with the spectrophotometer is general scanning and quantitation software that enables the user to program desired absorbance wavelengths for DNA quantitation.

Software Requirements

This data management system requires two types of software: commercial or "off the shelf" programs, and customized programs written by us expressly for use with the database software or for software used by individual modules.

The commercial software required includes a relational database program, backup software (usually provided with the backup drive), a pedigree drawing program, and optional network software. The relational database software is the keystone, providing storage, querying, and reporting capabilities. The relational database software must also allow each module to access data both within and between modules on different computers. One of the most useful features of this database is the ability to output data into a pedigree drawing format, so the software chosen for the pedigree drawing should be able to import files. We use Paradox version 5.0 (Borland Intl., Scotts Valley, CA) for the relational database, Colorado Backup version 2.80, Cyrillic 2.1 (Cherwell Scientific, Oxford, UK) for the pedigree drawing, and Windows for Workgroups 3.11 on the server computer. Computers used for the individual modules use either Windows for Workgroups 3.11 or Windows NT 3.51/4.0.

The customized programs come in three types. The first type facilitates the use of the database software. These programs consist of data entry forms, reports, and queries that are part of the relational database program options; only knowledge of that software is needed. The second type of customized program builds on the options within the database software (for Paradox these programs are written in ObjectPal). For example, we have created "smart forms" that aid data entry by filling some fields with prespecified default values, skipping inappropriate fields based on previously entered values (e.g., skipping an IF YES, SPECIFY field when "no" was entered into the previous field), and automatically calculating sums and differences. Other forms provide a "button" that when "pressed" executes other programs, e.g., backing up the data files or archiving old data. These custom programs, which modify and extend features within the database software, call for additional knowledge of that software.

The primary output for the DNA database is a report listing the box location, current volume, and total amount of DNA per vial and per subject. Other reports can be generated that essentially function as flags. For example, reports are generated when a particular DNA vial is of poor quality (e.g., out-of-range 260/280-nm ratio) or when the amount of DNA for a particular subject goes below a user-specified value, thereby alerting the technician to begin cell culture for the extraction of new DNA. The main output file can also be interfaced with the cell line, clinical, and genotype databases.

Cell Line Module

Structure. The cell line module consists of three related tables (Fig. 4). The *Growth* table contains data about each growth attempt, recording the number of attempts, quality of the growth, and reasons for any failure. The *Storage* table contains the box and coordinate location data for each cell line vial in the freezer, giving an up-to-date inventory of cell lines available for each subject studied. The *Usage History* table records any additions or removals to the freezer, thus tracking the usage of cell lines, and facilitating error checks and audits.

Data flow. When a blood sample arrives at the laboratory, an attempt is made to grow a cell line. The vigor and quality of the culture and relevant dates are recorded in the *Growth* table for every growth attempt. Once a cell line is grown successfully, the storage information is entered into the *Usage History* table, with a field showing that it is a new addition. These data are automatically copied into the *Storage* tables. When a

cell line is removed from the freezer, this fact is also entered into the *Usage History* table, with the field showing it as a new removal. The vial is then automatically deleted from the *Storage* table. As a result, the *Storage* table always contains an accurate inventory of the available cell lines and their locations. The *Usage History* table contains a record of all additions and removals, and thus acts as an archive for the *Storage* table.

Reports are generated to identify subjects needing to have a blood sample redrawn because the culture failed, to summarize the number and locations of cell line vials stored for each individual, and to alert laboratory staff that the supply of cell line vials for an individual has gone below a user-specified value.

Genotype Module

Structure. The genotype module consists of four related tables (Fig. 5). The *Genotype* table contains the marker genotypes for each subject, in the form of arbitrary allele numbers. The exact allele size in base pairs corresponding to each arbitrary allele number is recorded in the *Allele Size* table. The *Reader* table contains the information on who read the genotypes in each family, with MARKER ID specified. *Allele Size* and *Reader* are linked to other tables via FAMILY ID, since arbitrary allele numbers are assigned within each family. The last table, *Markers*, contains the reference information for all markers, linking the MARKER ID to the marker name(s), chromosome, and location (if desired, multiple *Markers* tables can be used to group markers in a convenient manner, such as by chromosome).

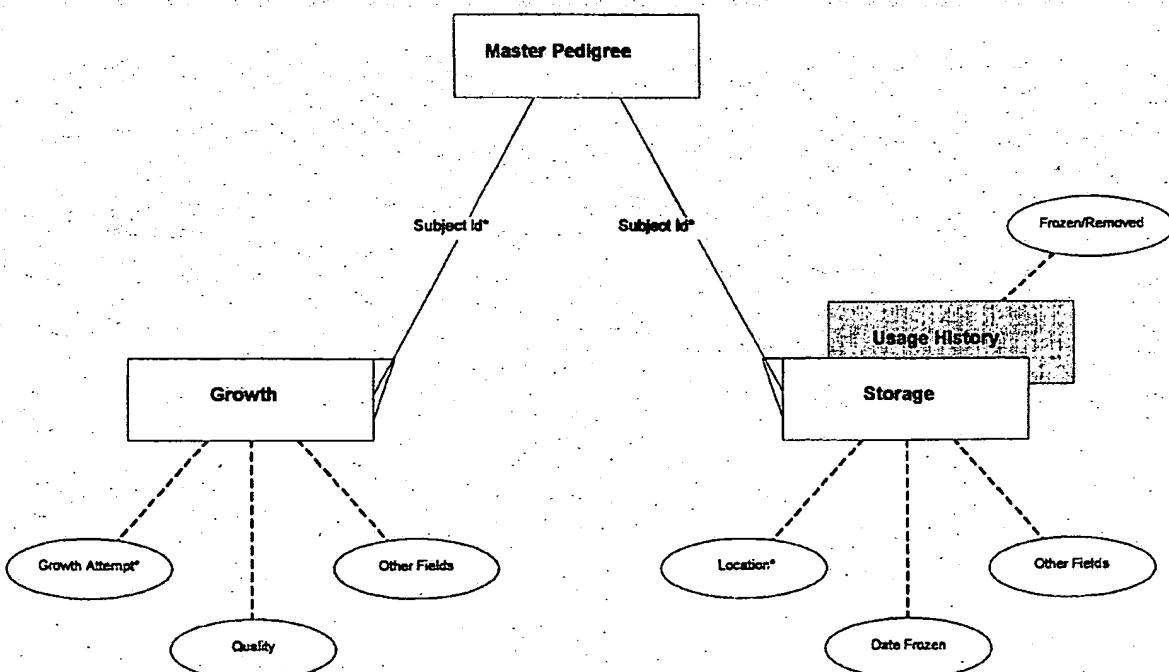


Fig. 4. The cell line data module tracks the growth and storage of each vial of lymphoblastoid cells generated for each subject. See Figure 2 legend for explanation of symbols.

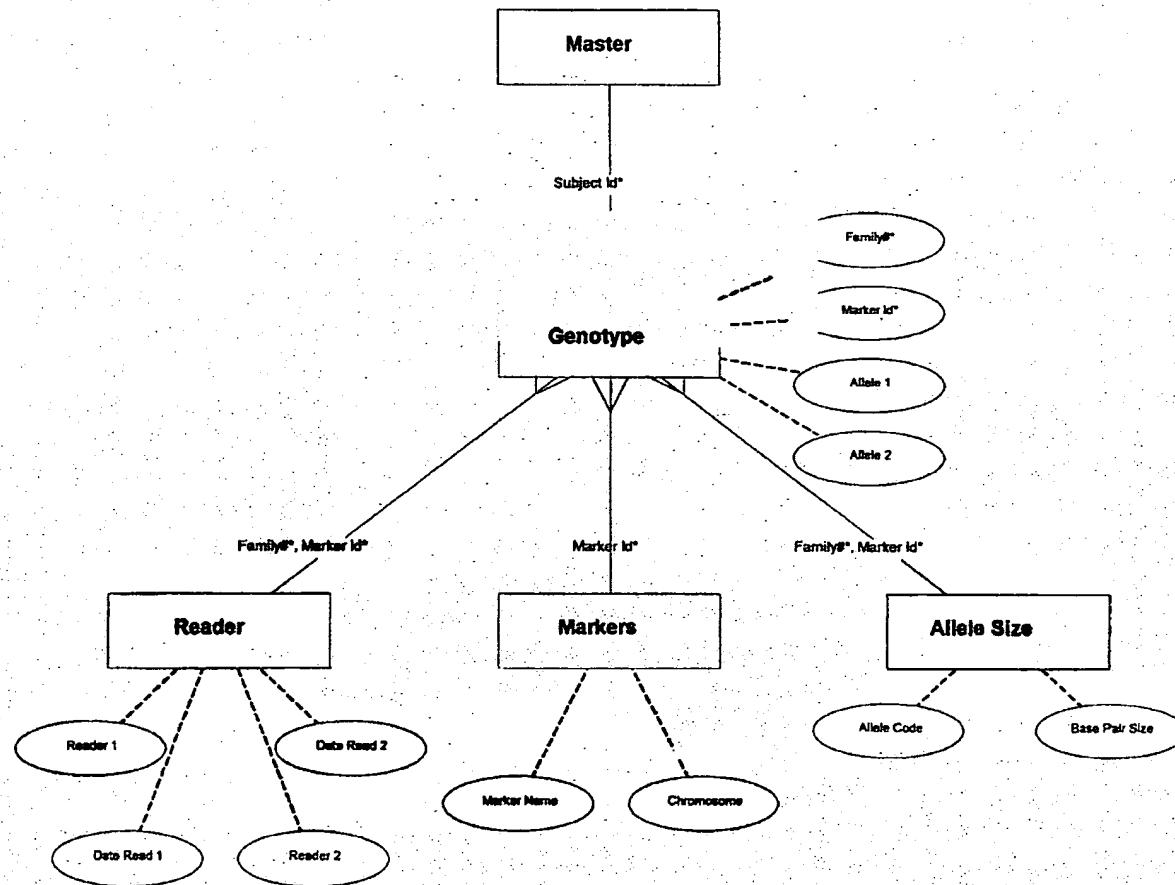


Fig. 5. The genotype data module tracks the results of the polymorphic DNA marker analyses for each subject as well as descriptive data about the markers used and their chromosomal locations. See Figure 2 legend for explanation of symbols.

Data flow. After a family has been clinically evaluated, the individuals with a DNA sample who are required for linkage analysis are selected for genotyping. This is noted in the *Master Pedigree* table, in a field called **GENOTYPING**. Another field, **NEED FOR PEDIGREE**, designates the individuals selected for genotyping along with individuals required to connect the pedigree structure, e.g., parents of a sib-pair.

Once the genotypes for each marker are determined, they are entered into the *Genotype* table. If the genotypes were read automatically, these data are set in an importable format by a semiautomated routine. Output from the ABI 373 sequencer is binned using the program Genetic Analysis System (GAS 2.0; GAS © Alan Young, Oxford University, 1993–1995), whose text output is imported into the database tables. If genotypes are read manually, the information is entered directly into the *Genotype* table. This is accomplished with a form that requires entering **MARKER ID** only once per family and simplifies data entry by allowing entry into *Genotype*, *Reader*, and *Allele Size* tables all at once. This guarantees complete data for every genotype, e.g., that every arbitrary allele number corresponds to an absolute allele size value in the database.

The primary output from the genotype module is linked with the phenotype data from the clinical module to generate linkage files for analysis. Other outputs can also be generated, e.g., status reports summarizing genotype progress by individual or marker.

Pedigree Reports

The most useful report format for family studies is often the pedigree itself. Therefore, we developed a method for importing any data of interest from the database into a pedigree drawing program, where the data are displayed directly on the pedigree. This approach preserves the flexibility of the relational database while displaying the data in the way that is most intuitive for genetic researchers.

For each report format, a program collects the data of interest from the relevant tables in the database, joins these data with the pedigree structure information in the *Master Pedigree* table, and formats the joined data for importing into the pedigree drawing program. Residual errors in pedigree structure are easily detected at this point, since any errors will cause the pedigree drawing program to either reject the import file or

Hidden linkage: a comparison of the affected sib pair (ASP) test and transmission/disequilibrium test (TDT)

R. E. McGINNIS

Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia,
PA 19104-6145

(Received 3.11.97. Accepted 16.3.98)

SUMMARY

I compare the transmission/disequilibrium test (TDT) and affected sib pair (ASP) test under a general algebraic model describing a bi-allelic disease locus. Assuming linkage to a bi-allelic marker, I derive two binomial probabilities, one for parental allele 'transmission' (P_t) which determines the magnitude of the TDT χ^2 statistic (χ^2_{TDT}), and a second for identity-by-descent (ibd) marker allele 'sharing' (P_s) which determines the magnitude of the ASP test statistic (χ^2_{ASP}). I also consider the ASP test applied to a completely polymorphic marker and demonstrate that the probability of ASP marker allele sharing (P_s) is identical to P_s observed for a bi-allelic marker in equilibrium with the disease locus. I present a general framework for determining the power of the TDT and ASP test based on expressions for P_t , P_s and the proportion (H/F) of ascertained parents who are informative at the marker. Two previous analytic investigations of TDT power based on the work of Ott (1989), and Risch & Merikangas (1996) are shown to be special cases of this general framework. In addition, I show the relationship between the framework I present and a third analytic investigation of TDT power for multi-allelic markers based on the work of Sham & Curtis (1995).

INTRODUCTION

Linkage has been demonstrated between insulin-dependent diabetes mellitus (IDDM) and the insulin gene region on chromosome 11p15.5 on the basis of linkage analysis by the transmission/disequilibrium test or TDT (McGinnis *et al.* 1991; Spielman *et al.* 1993). Linkage was demonstrated at the insulin 5'VNTR, a hypervariable marker that is extremely polymorphic, but whose VNTR alleles fall into two main size classes in Caucasians, thus forming a natural bi-allelic (+/−) marker. The + alleles were discovered to be positively associated with IDDM in case-control studies (Bell *et al.* 1984). Subsequent studies then demonstrated linkage in families collected for Genetic Analysis Workshop 5 (GAW5) by TDT analysis of GAW5 parents who were heterozygous (+/−) under the 5'VNTR bi-allelic categories (Spielman *et al.* 1993; see also Thomson *et al.* 1989, Julier *et al.* 1991).

The very strong evidence for linkage provided by the TDT ($\chi^2 = 8.28$, $p < 0.005$) was both surprising and puzzling because identity-by-descent (ibd) sharing of 5'VNTR alleles in affected sib pairs (ASPs) yielded no evidence for linkage in the same GAW5 families. Indeed, evidence for linkage was completely undetected or 'hidden' because the proportion of alleles shared by ASPs did not exceed the null hypothesis value of 0.5 in two different types of ASP analysis. On one hand, there was no increase in ASP allele sharing when the analysis included all GAW5 families in which both parents were informative for any two lengths of 5'VNTR allele (Spielman *et al.* 1989; Cox & Spielman, 1989). On the other hand, when the analysis included only those ASP parents who were evaluated by the TDT, namely those heterozygous (+/−) when the 5'VNTR is con-

Address for correspondence: Dr. Ralph McGinnis, Senior Investigator, SmithKline Beecham, New Frontiers Science Park (North), Third Avenue, Harlow, Essex CM19 5AW.

Comparison of the TDT and ASP test

calculated for a marker that is distinct from the disease locus. Analysis of the equations shows that TDT power is greatly increased if disequilibrium is strong and if the disease allele and positively associated marker allele have similar population frequencies. The equations also show that the superior power of the TDT compared to the ASP test is greatest when susceptibility loci confer modest disease risk, as indicated by low values of the penetrance ratio r . When a marker is strongly associated with a disease locus that contributes modest disease risk, $|P_t - 0.5| \gg (P_s - 0.5) \approx 0$. Thus, the TDT is likely to play an important role in detecting and replicating linkages to loci responsible for complex genetic disease.

I am deeply grateful to Richard Spielman for encouragement and valuable suggestions as this work developed. I am also indebted to Warren Ewens for valuable comments and for criticism that improved the manuscript. This research was supported by NIH grants DK46618 and DK47481 and by grant 193189 from the Juvenile Diabetes Foundation.

REFERENCES

BELL, G. I., HORITA, S. & KARAM, J. H. (1984) A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 33, 176-183.

BLACKWELDER, W. C. & ELSTON, R. C. (1985) A comparison of sib-pair linkage tests for susceptibility loci. *Genet. Epidemiol.* 2, 85-97.

CLERGET-DARPOUX, F., BABRON, M. C. & BICKEBÖLLER, H. (1995) Comparing the power of linkage detection by the transmission disequilibrium test and identity-by-descent test. *Genet. Epidemiol.* 12, 583-588.

COX, N. J. & SPIELMAN, R. S. (1989) The insulin gene and susceptibility to IDDM. *Genet Epidemiol* 6, 65-69.

EWENS, W. J. & SPIELMAN, R. S. (1997) Disease associations and the transmission/disequilibrium test (TDT). *Current Protocols in Human Genetics* 1.12.1-1.12.13.

★ JULIER, C., HYER, R. N., DAVIES, J., MERLIN, F., SOULARUE, P., BRIANT, L., CATHELINEAU, G., et al. (1991) Insulin-IGF2 region on chromosome 11p encodes a gene implicated in HLA-DR4-dependent diabetes susceptibility. *Nature* 354, 155-159.

KAPLAN, N. L., MARTIN, E. R. & WEIR, B. S. (1997) Power studies of the transmission/disequilibrium tests with multiple alleles. *Am. J. Hum. Genet.* 60, 691-702.

MCGINNIS, R. E., SPIELMAN, R. S. & EWENS, W. J. (1991) Linkage between the insulin gene (IG) region and susceptibility to insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet. Suppl.* 49, A476.

MÜLLER-MYHSOK, B. & ABEL, L. (1997) Genetic analysis of complex diseases. *Science* 275, 1328-1329.

OTT, J. (1989) Statistical properties of the haplotype relative risk. *Genet. Epidemiol.* 6, 127-130.

OTT, J. (ed) (1991) *Analysis of human genetic linkage*. Johns Hopkins University Press, Baltimore.

RISCH, N. (1990) Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* 46, 229-241.

RISCH, N. & MERIKANGAS, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516-1517.

PEARSON, E. S. & HARTLEY, H. O. (ed) (1954) *Biometrika tables for statisticians*. Vol 1. Cambridge University Press.

SCHAID, D. J. & SOMMER, S. S. (1994) Comparison of statistics for candidate-gene association studies using cases and parents. *Am. J. Hum. Genet.* 55, 402-409.

SHAM, P. C. & CURTIS, D. (1995) An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann. Hum. Genet.* 59, 323-336.

SPIELMAN, R. S., BAUR, M. P. & CLERGET-DARPOUX, F. (1989) Genetic analysis of IDDM: summary of GAW5 IDDM results. *Genet. Epidemiol.* 6, 43-58.

★ SPIELMAN, R. S., MCGINNIS, R. E. & EWENS, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52, 506-516.

SPIELMAN, R. S. & EWENS, W. J. (1996) The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* 59, 983-989.

TERWILLIGER, J. D. & OTT, J. (1992) A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum. Hered.* 42, 337-348.

★ THOMSON, G., ROBINSON, W. P., KUHNER, M. K. & JOE, S. (1989) HLA, insulin gene, and Gm associations with IDDM. *Genet. Epidemiol.* 6, 155-160.

WEIR, B. S. (ed) (1996) *Genetic data analysis II*, 2nd ed., Sinauer Associates Inc., Sunderland, MA.

APPENDIX I

Derivation of expressions for P_s , P_t , H

The derivations assume the general model of a bi-allelic marker and linked bi-allelic disease locus that is the only locus that underlies disease susceptibility (see General algebraic model of linkage in the main text). I begin the derivation of P_s and P_t (equations (1) and (2)) by first deriving

Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-dependent Diabetes Mellitus (IDDM)

Richard S. Spielman,* Ralph E. McGinnis,* and Warren J. Ewens†

*Department of Genetics, University of Pennsylvania School of Medicine, and †Department of Biology, University of Pennsylvania, Philadelphia

Summary

A population association has consistently been observed between insulin-dependent diabetes mellitus (IDDM) and the "class 1" alleles of the region of tandem-repeat DNA (5' flanking polymorphism [5'FP]) adjacent to the insulin gene on chromosome 11p. This finding suggests that the insulin gene region contains a gene or genes contributing to IDDM susceptibility. However, several studies that have sought to show linkage with IDDM by testing for cosegregation in affected sib pairs have failed to find evidence for linkage. As means for identifying genes for complex diseases, both the association and the affected-sib-pairs approaches have limitations. It is well known that population association between a disease and a genetic marker can arise as an artifact of population structure, even in the absence of linkage. On the other hand, linkage studies with modest numbers of affected sib pairs may fail to detect linkage, especially if there is linkage heterogeneity. We consider an alternative method to test for linkage with a genetic marker when population association has been found. Using data from families with at least one affected child, we evaluate the transmission of the associated marker allele from a heterozygous parent to an affected offspring. This approach has been used by several investigators, but the statistical properties of the method as a test for linkage have not been investigated. In the present paper we describe the statistical basis for this "transmission test for linkage disequilibrium" (transmission/disequilibrium test [TDT]). We then show the relationship of this test to tests of cosegregation that are based on the proportion of haplotypes or genes identical by descent in affected sibs. The TDT provides strong evidence for linkage between the 5'FP and susceptibility to IDDM. The conclusions from this analysis apply in general to the study of disease associations, where genetic markers are usually closely linked to candidate genes. When a disease is found to be associated with such a marker, the TDT may detect linkage even when haplotype-sharing tests do not.

Introduction

A crucial first step in finding gene loci that contribute to a genetic disease is to demonstrate linkage with a gene or DNA sequence of known location (a "marker," usually a DNA polymorphism). A number of investigators have used this approach in the study of diabetes

mellitus. Bell et al. (1984) described a population association between insulin-dependent diabetes mellitus (IDDM) and the 5' flanking polymorphism (5'FP), an RFLP adjacent to the insulin gene on chromosome 11p. Although it is not clear that insulin or the insulin gene itself plays a role in the pathogenesis of IDDM, the association has been found consistently in population studies (for a summary, see Cox et al. 1988). In unaffected controls, the frequency of the smaller, or "class 1," alleles is approximately .70-.75, while in IDDM patients the frequency is somewhat higher: .80-.85. This finding provides *indirect* evidence for linkage between the insulin gene region and genes that influence

Received July 13, 1992; revision received November 10, 1992.

Address for correspondence and reprints: Dr. Richard Spielman, Department of Genetics, University of Pennsylvania School of Medicine, 422 Curie Boulevard, Philadelphia, PA 19104-6145.

© 1993 by The American Society of Human Genetics. All rights reserved.
0021-9297/93/5203-0008\$02.00

susceptibility to IDDM, since an association between disease and marker may be due to disequilibrium between linked loci. However, the problem with inferring linkage from population association is that association can occur in the absence of linkage—for example, as a result of population stratification. Thus it is not valid to use the presence of association as a test for linkage if population stratification is a possibility.

For this reason, tests of linkage that do *not* depend on association were carried out by various investigators. In most of these studies, there was no direct evidence for linkage (Hitman et al. 1985; Ferns et al. 1986). In larger samples, the distribution of S'FP alleles in 33 affected sib pairs (ASPs) with IDDM (Cox et al. 1988) or in the 95 ASPs studied in Genetic Analysis Workshop 5 (GAW5) (Cox and Spielman 1989; Spielman et al. 1989) failed entirely to provide evidence for linkage. Thus the absence of cosegregation within families suggested that the association was due to population stratification rather than to disequilibrium with a linked locus.

However, other approaches have suggested that the association is not due solely to stratification. Using the method of Field et al. (1986), Thomson et al. (1989) analyzed the GAW5 family data by the following method. In each family, the four parental S'FP alleles were assigned to one of two categories: (1) transmitted to at least one diabetic offspring ("diseased") and (2) not transmitted to any affected offspring ("control"). This method has been termed "AFBAC," for "affected family-based controls" (Thomson 1988). As tested by a conventional χ^2 , the frequency of S'FP class 1 alleles in the diseased category (.83) was significantly higher than that in the controls (.69) ($p < .01$). Since the control and disease samples are obtained from the same individuals, the contribution of stratification to the association is reduced or eliminated. However, the comparison does not provide a *direct* test for linkage.

In the present paper we describe a procedure which tests directly for linkage between a disease and marker locus which shows population association; this test is not affected by the presence of stratification. The data for the test are from families with one or more affected offspring and at least one parent who is *heterozygous* for a marker allele (e.g., S'FP class 1) associated with the disease. The test procedure compares (a) the number of times that such heterozygous parents transmit the associated marker to an affected offspring with (b) the number of times that they transmit the alternate marker allele. Because of this focus on alleles transmitted to

affected offspring, the test shares some features with the concept of haplotype relative risk (HRR; Falk and Rubinstein 1987) and with the AFBAC test of association (Field et al. 1986; Thomson et al. 1989; Field 1991) described above. However, because our emphasis is on testing for linkage, the actual tests are different. Since GAW5 (Spielman et al. 1989), the principle underlying this linkage test has been used explicitly (McGinnis et al. 1991) or implicitly (Owerbach et al. 1990; Julier et al. 1991) in other investigations, to provide additional evidence that determinants of IDDM are located in the insulin gene region.

In GAW5, Ott presented the formal theory which is necessary for any test of a hypothesis based on a comparison of frequencies of marker alleles transmitted or not transmitted to affected offspring. His analysis showed that the probabilities of the various possible combinations of transmitted and nontransmitted marker locus alleles are determined by the association (disequilibrium) parameter δ and the recombination fraction θ between the loci. However, we show below that the χ^2 procedure used as a test of association (i.e., AFBAC) is not, in general, valid as a test of linkage, and we derive a procedure which is valid. We also show (1) that our testing procedure also provides a test for association between the two loci (indeed, the test can detect linkage only if association exists); (2) the relationship of this test to tests based on sharing of haplotypes or genes (identical by descent) in ASPs, affected sib trios, etc.; and (3) the result of applying this test to data on the S'FP in IDDM.

The Transmission Test for Linkage Disequilibrium

The transmission/disequilibrium test (TDT) considers parents who are heterozygous for an allele associated with disease and evaluates the frequency with which that allele or its alternate is transmitted to affected offspring. Compared with conventional tests for linkage, the TDT has the advantage that it does not require data either on multiple affected family members or on unaffected sibs. However, the TDT has the disadvantage that it can detect linkage between the marker locus and the disease locus only if association (due to linkage *disequilibrium*) is present.

In the following sections we describe the properties of the TDT as a test of significance for linkage. We then discuss the relationship of the TDT to tests of linkage that are based on shared haplotypes in ASPs.

We assume a disease locus D , with disease allele D_1 ,

and haplotype-sharing χ^2 's can be used, separately or together, to test for linkage. These considerations also generalize to sibships with four or more affected.

We have shown above how a χ^2 statistic to test transmission/disequilibrium can be calculated for data in which all families have the same number of affected children. In any real set of data, we can expect to observe families with varying numbers of affected children. In such a case we recommend simply combining all affected children in the data, irrespective of number of affected in the family, in one overall transmission/disequilibrium χ^2 statistic of the form $(B-C)^2/(B+C)$, where B is the total number of transmissions of M_1 to affected children and C is the total number of transmissions of M_2 . In the case where segregation distortion at the M locus is a possibility, an aggregate 2×2 -table χ^2 is appropriate, corresponding to that discussed above for the case of one affected child per family. We use such a χ^2 procedure below in the Results subsection.

Data and Results

Data

The data for this study were assembled for GAW5 from 94 families with two or more IDDM children (Baur et al. 1989; Spielman et al. 1989). For GAW5, Southern blots of genomic DNA digested with *Pvu*II were hybridized with phins 310 (Bell et al. 1984), a probe specific for the S'FP, and alleles were assigned by eye to one of three classes corresponding to fragment size. (Class 1 is smallest, and class 3 is largest.) Gel positions of genomic bands and markers were also recorded; for the present reanalysis we assigned restriction fragments to allele class 1 if they were smaller than 1 kb, to class 2 if they were 1–2 kb, and to class 3 if they were larger than 2 kb. Since our analysis focuses on the role of class 1 alleles, class 2 and class 3 alleles were grouped together as class X. Among the 94 families, there were 53 in which at least one parent was heterozygous for class 1 and class X alleles.

Results

In order to demonstrate the usefulness of the TDT, we review the findings with respect to population association and haplotype sharing. The family data obtained for GAW5 do not lend themselves to a conventional association study, which would include unrelated controls. However, when just unrelated diabetics (the oldest affected sib in each family) are considered, the frequency of class 1 alleles in the present

Table 5

TDT for Alleles 1 and X of S'FP in IDDM: Data for 1/X Parents of All Affected Children

	NO. OF ALLELES TRANSMITTED			SIGNIFICANCE (p)	
	1	X	Total	χ^2_{ad}	
Observed	78	46	124	8.26	.004
Expected	62	62			

(GAW5) data is $138/162 = .85$. This value is similar to those reported for "random" diabetics and is higher than that found in unrelated controls, as has been observed elsewhere (Cox et al. 1988).

An analysis of haplotype sharing in the GAW5 family data was previously carried out by Cox and Spielman (1989). Using the χ^2 test statistic of equation (12) ("Y" of Blackwelder and Elston [1985], applied strictly to ASPs), Cox and Spielman (1989) did not find even modest departures from random sharing. This result also held when they considered only families with at least one parent heterozygous for class 1/class 3 at the S'FP. (For the corresponding test by equation [7] or t_2 of Blackwelder and Elston [1985], see table 7 below.) Thus there is population association but no evidence for linkage, by conventional tests.

However, when linkage is tested by the TDT, a different conclusion emerges (table 5). There were 57 parents heterozygous for alleles 1 and X of the S'FP; these parents transmitted 124 alleles (78 class 1 alleles and 46 class X alleles) to their diabetic offspring. Under the hypothesis of no linkage, the expected number of transmissions of 1 and X is equal (i.e., 62). When equation (5) is used, the difference observed is highly significant; $\chi^2_{\text{ad}} = (78-46)^2/124 = 8.26$, $p = .004$.

As explained above, the difference found with the TDT could be due to an "artifact" of meiotic segregation distortion, which would be expected to apply to both affected and unaffected offspring, if unrelated to disease. For this reason, we compared affected and unaffected offspring with respect to transmitted class 1 and class X alleles. The results are shown in table 6. Among affected offspring, 78 (63%) of 124 alleles received from heterozygous parents were class 1. The corresponding figure for unaffected offspring was 42 (40%) of 104; the difference is highly significant ($\chi^2_1 = 11.5$, $p < .001$). This result confirms the finding of linkage; there is no evidence for segregation distortion.

Table 6

Comparison of Alleles 1 and X of 5'FP Transmitted to IDDM-affected Offspring and Unaffected Sibs

	NO. OF ALLELES TRANSMITTED			SIGNIFICANCE	
	1	X	Total	χ^2_{dd}	(p)
Affected	78	46	124	11.5	<.001
Unaffected	42	62	104		

NOTE.—Data for 1/X parents.

The strong evidence for linkage, based on the TDT, stands in striking contrast to conclusions obtained, in earlier studies, from the Y statistic for haplotype sharing. However, the TDT (above) and the Y statistic were based on overlapping but not identical sets of families. This discrepancy arose because families with only one parent heterozygous 1/X could be used for the TDT but not for the Y statistic. Furthermore, parents with two distinguishable class 1 alleles were used for Y but not for the TDT.

These differences in the data led us to ask the following question: Is the failure to find linkage with the Y statistic due entirely to the difference between the samples used, or are linkage tests based on haplotype sharing inherently less sensitive than the TDT for the present data? To answer this question, we applied the transmission/disequilibrium (χ^2_{dd}) and haplotype-sharing (χ^2_{hs}) tests to exactly the same data. Not all the data from table 5 can be used, because some are from simplex families or from sibships with more than two affected sibs. Accordingly, we used just those families with at least one 1/X parent and exactly two affected sibs, as appropriate for equations (4)–(8). Table 7 shows the data in the form of definitions (4).

For the TDT we compare i with j , by equation (6), and obtain $\chi^2_{dd} = 3.60$ ($p = .058$). Unlike the corresponding test above ($\chi^2_{dd} = 8.26$), the present comparison is not "quite" significant. Although the proportion of class 1 alleles transmitted here ($54/90 = .60$) is almost the same as that in table 5 ($78/124 = .63$), the χ^2_{dd} is smaller, and the significance level is less striking, because of the smaller sample size.

For the haplotype-sharing test, we compare (a) the number of parents ($i+j = 21$) who transmitted the same allele (1 or X) to both affected children with (b) the number ($b-i-j = 24$) who transmitted different alleles.

This is equivalent to comparing the number of ASPs who received the same allele ("shared") with the number who received different alleles ("unshared"). The resulting χ^2_{hs} (0.20) is not significant, and the difference is in the opposite direction of that predicted by linkage, presumably reflecting random variation. There is not even a "trend" toward increased sharing.

Thus, in the present analysis of a single body of data, we see the discrepancy identified in earlier reports. There is a population association between IDDM and the class 1 allele of the 5'FP, but sharing of alleles by affected sibs (cosegregation) provides no evidence for linkage. Nevertheless, there is highly significant evidence of linkage in the TDT.

Discussion

Linkage studies for so-called complex genetic diseases pose problems not found in standard linkage analysis. Because these diseases, in general, have reduced penetrance, unaffected family members usually provide much less information for linkage than do affected members. In this situation, it is essential to study families with multiple affected members and to focus on affected relatives, such as ASPs. The ASP approach has been applied with great success to unravel the role of the HLA complex in several diseases to which HLA appears to make a large contribution. For a locus that makes a modest contribution, however, the approach is severely limited. It has been shown by computer simulation (Cox and Spielman 1989) that, when ASPs are used, the power to detect linkage to such a locus is very modest and may require hundreds of ASPs. Furthermore, an additional consequence of the low penetrance

Table 7

Transmission from 45 1/X Parents of IDDM-affected Sib Pairs

	NO. OF 1/X PARENTS WHO TRANSMIT			
	Class 1 to Both Children	Class 1 to One Child and Class X to the Other	Class X to Both Children	TOTAL
$i = 15$	$b-i-j = 24$	$j = 6$	$b = 45$	

NOTE.—Data are for comparison of χ^2_{dd} (TDT) and χ^2_{hs} (haplotype sharing).

HLA DISEASE ASSOCIATIONS: Models for Insulin Dependent Diabetes Mellitus and the Study of Complex Human Genetic Disorders

Glenys Thomson

Department of Genetics, University of California, Berkeley, California 94720

CONTENTS

INTRODUCTION	31
THE HLA REGION	32
DISEASE ASSOCIATIONS WITH THE HLA SYSTEM	35
MECHANISMS OF DISEASE PREDISPOSITION	37
THEORETICAL ASPECTS OF HLA-DISEASE ASSOCIATIONS	39
INSULIN DEPENDENT DIABETES MELLITUS	43
PROSPECTS	46

INTRODUCTION

The genes of the human leukocyte antigen (HLA) region control a variety of functions involved in the immune response, and they influence susceptibility to over 40 diseases. Our understanding of the structure and function of the HLA genes, their disease associations, and the evolutionary features of this multigene family has benefited from recent advances in molecular biology, immunology, disease modeling, and population genetics. Although a great deal has been learned about the etiology of HLA-associated diseases such as insulin-dependent diabetes mellitus (IDDM) and rheumatoid arthritis (RA), the progress in disease studies in general has been slower than was initially

31

0066-4197/88/1215-0031\$0.20/0

Table 1 HLA disease associations. Data adapted from Tiwari and Terasaki (76) (OR values are combined estimates from a number of studies and cannot be directly calculated from Table), and IDDM DR data from Thomson et al (75).

Disease	Race ^a	Patients		Odds Ratio
		(% positive)	Controls (%)	
Ankylosing Spondylitis (AS)				
B27	C	89	9	69.1
B27	O	85	15	207.9
B27	N	58	4	54.4
Idiopathic Hemochromatosis				
A3	C	72	28	6.7
B7	C	48	26	2.9
B14	C	19	6	2.7
Insulin Dependent Diabetes Mellitus (IDDM)				
B8	C	40	21	2.5
B15	C	22	14	2.1
DR3	C	52	22	3.8
DR4	C	74	24	9.0
DR2	C	4	29	0.1
Rheumatoid Arthritis (RA)				
DR4	C	68	25	3.8
Celiac Disease (CD)				
B8	C	68	22	7.6
DR3	C	79	22	11.6
DR7	C	60	15	7.7
Multiple Sclerosis (MS)				
B7	C	37	24	1.8
DR2	C	51	27	2.7
Narcolepsy				
DR2	C	100	22	129.8
DR2	O	100	34	358.1

^aC = Caucasian O = Oriental N = Negro

antigen. With more recent typings of the class II loci, initially HLA-DR and -D, and now HLA-DP and -DQ, a number of other striking associations have been found. For example, 93% of patients with IDDM have DR3 or DR4 compared to 43% of controls, and 79% of Caucasian patients with celiac disease have DR3 compared to 22% of controls (76). For many diseases the DR (or other class II) antigens seem to be more strongly associated

The affected sib method is often more powerful in detecting the presence of disease predisposing genes than are standard association studies. Association studies require the existence of linkage disequilibrium between the alleles of the marker and disease predisposing loci. Significant linkage disequilibrium values are not usually expected for loci with recombination distances greater than approximately 2% (22). In contrast, deviations from random segregation of haplotype sharing values in affected sibs can be detected over much larger recombination distances between the marker loci and the disease predisposing loci (46). These do not require linkage disequilibrium for the detection of disease predisposing genes. In many instances the affected sib pair method has been a powerful statistical test for linkage; for example, the existence of HLA linked disease susceptibility to IDDM was statistically demonstrated using 15 affected sib pairs (9). However, in other instances for example, RA (45), demonstration of deviations of haplotype sharing from random expectations has required large sample sizes. Or the deviation may not have been demonstrated, even when a population association with the disease exists, for example, the association of the polymorphic region 5' to the insulin gene (5'FP) with IDDM (7, 74), implying either a high frequency of the disease predisposing allele (72) or the occurrence of sporadic cases of the disease. The affected sib pair method has also been generalized to consider the deviations of identity by state values, rather than identity by descent values from random expectations (31); this makes the method generally applicable for any polymorphic generic region, since all four parental chromosomes then do not have to be distinguishable.

A new approach to association studies avoids the use of a separate control population with its inherent problems of ascertainment bias and possible ethnic mismatching. This is to consider family studies where it is assumed the proband is an affected child, and full genetic information is available on both parents and all children (13, 14, 74). We term this method AFBA^C (Affected Family BA^Cted Controls). Within a family, each allele of the four from the parents is defined as belonging to the diseased category if it never appears in an affected individual. In this case, the alleles designated to the "nondiseased" category provide an appropriate control population with which to compare the "diseased" population for association studies. Application of this method to the GAW5 IDDM data set (74) confirmed the association of the 5'FP of the insulin gene (see 2).

It became clear early in the development of mathematical methods that single locus dominant or recessive models, with incomplete penetrance, were not sufficient to explain the inheritance patterns of the HLA associated diseases. Disease heterogeneity in the HLA region was implicated with demonstrations of synergistic effects for some diseases, notably DR3/DR4

memory system^{21,22}, the unique neurons described here could serve as memory storage elements, also activated in the retrieval process. □

Received 14 June; accepted 24 September 1991.

1. Wechsler, D. *Wechsler Memory Scale-Revised* (The Psychological Corporation, Harcourt Brace Jovanovich, San Antonio, 1987).
2. Meyer, V. & Yates, A. J. *J. Neurol. Neurosurg. Psychiat.* 18, 44–52 (1955).
3. Maher, B. *Brain Mechanisms Underlying Speech and Language* 122–145 (Grune & Stratton, New York, 1967).
4. Jones, M. K. *Neuropsychologia* 12, 21–30 (1974).
5. Petrides, M. *Neuropsychologia* 23, 601–614 (1985).
6. Goldstein, E. H., Canavan, A. G. M. & Pokey, C. E. *Cortex* 24, 41–52 (1988).
7. Murray, E. A., Gaffan, D. & Mishkin, M. *Soc. Neurosci. Abstr.* 14, 2 (1988).
8. Miyashita, Y. & Chang, H. S. *Nature* 331, 68–70 (1988).
9. Miyashita, Y. *Nature* 335, 817–820 (1988).
10. Snedecor, G. W. & Cochran, W. G. *Statistical Methods* 8th edn, 97 (Iowa State University Press, Ames, 1969).
11. Perret, D. I., Rolls, E. T. & Caan, W. *Expl. Brain Res.* 47, 329–342 (1982).
12. Desimone, R., Albright, T. D., Gross, C. G. & Bruce, C. J. *Neurosci.* 4, 2051–2062 (1984).
13. Schwartz, E. L., Desimone, R., Albright, T. D. & Gross, C. G. *Proc. natn. Acad. Sci. U.S.A.* 80, 5776–5778 (1983).
14. Arato, A. & Singer, W. *Nature* 330, 649–652 (1987).
15. Frégnac, Y., Shultz, D., Thorpe, S. & Blennerhast, E. *Nature* 333, 367–370 (1988).
16. Bruce, C. J. & Goldberg, M. E. J. *Neurophysiol.* 53, 603–635 (1985).
17. Maurer, K.-H. & Wise, S. P. *Expl. Brain Res.* 61, 229–244 (1988).
18. Furukoshi, S., Bruce, C. J. & Goldman-Rakic, P. S. J. *Neurophysiol.* 61, 331–349 (1989).
19. Milner, J. *Clin. Neurosurg.* 19, 422–446 (1972).
20. Squire, L. R., Cohen, N. J. & Nadel, L. *Memory Consolidation: Psychobiology of Cognition*, 185–210 (Lawrence Erlbaum, Hillsdale, 1984).
21. Insausti, R., Amaro, D. G. & Cowan, W. M. J. *Comp. Neural.* 284, 356–395 (1987).
22. Webster, M. J., Ungerleider, L. G. & Bachelder, J. J. *Neurosci.* 11, 1095–1116 (1991).
23. Zahn, C. T. & Roskies, R. Z. *IEEE Trans. Comput.* 21, 269–281 (1972).
24. Rolls, E. T. et al. *J. Neurosci.* 9, 1835–1845 (1989).
25. Gross, C. G., Rocha-Miranda, C. E. & Bender, D. B. J. *Neurophysiol.* 38, 96–111 (1972).

ACKNOWLEDGEMENTS. We thank Y. Hotta for encouragement. This work was supported by a Grant for Scientific Research on Priority Areas from the Japanese Ministry of Education, Science and Culture.

Insulin-IGF2 region on chromosome 11p encodes a gene implicated in HLA-DR4-dependent diabetes susceptibility

C. Julier*, R. N. Hyert, J. Davies*, F. Martin*, P. Soularue*, L. Briant†, G. Cathelineau§, I. Deschamps||, J. I. Rotter||, P. Froguel*§, C. Boitard||, J. I. Bell†* & G. M. Lathrop*

* Centre d'Etude du Polymorphisme Humain, 27 rue Juliette Dodu, 75010 Paris, France

† Molecular Immunology Group, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK

‡ INSERM U100, Hôpital Purpan, Place du Docteur Baylac, 31059 Toulouse, France

§ Service d'Endocrinologie, Hôpital St Louis, 2 Place du Docteur Fournier, 75010 Paris, France

|| INSERM U25, Hôpital Necker, 161 Rue de Sèvres, 75015 Paris, France

¶ Division of Medical Genetics, Cedars-Sinai Medical Center, and UCLA School of Medicine, Los Angeles, California 90048, USA

A CLASS of alleles at the VNTR (variable number of tandem repeat) locus in the 5' region of the insulin gene (*INS*) on chromosome 11p is associated with increased risk of insulin-dependent diabetes mellitus (IDDM)^{1–6}, but family studies have failed to demonstrate linkage^{5,7}. *INS* is thought to contribute to IDDM susceptibility but this view has been difficult to reconcile with the lack of linkage evidence^{8–10}. We thus investigated polymorphisms of *INS* and neighbouring loci in random diabetics, IDDM multiplex families and controls. HLA-DR4-positive diabetics showed

an increased risk associated with common variants at polymorphic sites in a 19-kilobase segment spanned by the 5' *INS* VNTR and the third intron of the gene for insulin-like growth factor II (*IGF2*). As *INS* is the major candidate gene from this region, diabetic and control sequences were compared to identify all *INS* polymorphisms that could contribute to disease susceptibility. In multiplex families the IDDM-associated alleles were transmitted preferentially to HLA-DR4-positive diabetic offspring from heterozygous parents. The effect was strongest in paternal meioses, suggesting a possible role for maternal imprinting. Our results strongly support the existence of a gene or genes affecting HLA-DR4 IDDM susceptibility which is located in a 19-kilobase region of *INS-IGF2*. Our results also suggest new ways to map susceptibility loci in other common diseases.

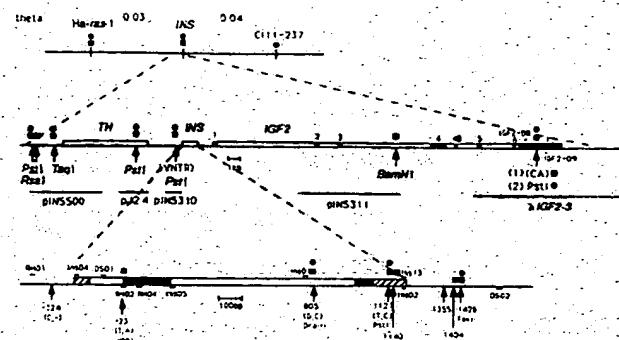


TABLE 3 IBD and segregation of alleles at *INS* locus

(a) Paternal and maternal IBD for diabetic offspring compared with probands in multiplex families

Variant	Alleles from father		Alleles from mother		Combined	
	1	0	1	0	1	0
Haplotype	68	54	50	72	118	126
5' VNTR	31	13†	17	15	48	28*
805/PstI	30	13†	12	14	42	27
1,127/PstI	29	10†	12	10	41	20†
1,428/FokI	30	10†	12	11	42	21†

(b) Segregation of paternal and maternal insulin alleles in relation to the HLA type of diabetic offspring in data from this study and GAW

HLA genotypes and *INS* allele transmitted to diabetic offspring from informative (+/-) meioses*

Variant	Parental origin	DRX/X		DR3/X, DR3/3		DR4/X, DR4/4		DR3/4		All DR4	
		+	-	+	-	+	-	+	-	+	-
1,127/PstI											
This study	father	2	3	7	11	22	8*	20	7*	42	15‡
	mother	0	0	2	7	7	6	15	14	22	20
	total	2	3	9	19	29	14*	35	21	64	35†
5' VNTR											
This study	father	2	3	8	13	22	9*	24	9*	46	18‡
	mother	0	0	6	8	11	9	20	17	31	26
	total	2	3	14	21	33	18*	44	26*	76	44*
GAW											
father	0	0	8	6	15	4*	17	7*	32	11‡	
mother	1	2	3	3	10	4	13	10	23	14	
total	1	2	11	9	25	8†	30	17	55	25‡	
Combined†											
father	2	3	15	17	36	12‡	37	15†	73	27‡	
mother	1	2	9	11	19	12	29	25	48	37	
total	3	5	24	28	55	24‡	66	40†	121	64§	

IBD statistics were obtained by counting the number of haplotypes shared by diabetic probands and their affected sibs. Overall, 29 patients shared two haplotypes identical to those of the proband, 60 shared one, and 33 shared none. Expectations for these classes without linkage are 30.5, 61 and 30.5, respectively. In a, where results are presented for meioses that were informative for each of the *INS* loci, *P*-values for the test of linkage were calculated from two-sided chi-squared tests. The tests are not independent because of linkage disequilibrium. Alleles of the 5' *VNTR* locus have been combined into size classes as described in Table 1. b, The alleles that have been transmitted to the affected offspring in informative meioses, after children were classified by their HLA type. Families from the GAW study that had recombination between *INS* region markers were deemed to contain genotype errors and were removed before analysis. The two studies included eight informative families in common which were removed from the GAW data before their combination. Contingency table analysis of the 5' *VNTR* data from both studies showed significant heterogeneity of the transmission frequencies by HLA genotype of the affected offspring ($P < 0.025$) and by parental sex ($P < 0.04$). *P*-values were calculated from two-sided chi-squared tests for deviations from random transmission of the *INS* allele in each HLA-DR group. * $P < 0.05$; † $P < 0.01$; ‡ $P < 0.001$; § $P < 10^{-4}$; ¶ $P < 10^{-5}$.

† After correction for families common to both studies.

‡ Plus, transmission of IDDM-associated allele; minus, transmission of another allele.

carried two 5' VNTR class 1 alleles but were homozygous at other *INS* sites. As before, we found no evidence of linkage when these meioses were included in the counts of haplotype sharing (Table 3a). But in those offspring whose parents were heterozygous for one or more of the IDDM-associated *INS* variants, the IBD probabilities were significantly greater than 50% at three of the four *INS* sites studied. Preferential transmission of 5' VNTR alleles was seen in meioses from parents who carried a single class 1 allele. Surprisingly, maternal IBD probabilities did not differ from 50%, whereas the paternal effect was significant at all *INS* sites characterized in the multiplex families (Table 3a). No preferential segregation was seen in nondiabetic offspring, or in offspring of reference families obtained from the Centre d'Etude du Polymorphisme Kumain (data not shown).

As the *INS* association was significant only in HLA-DR4-positive diabetics, we further subdivided meioses by the HLA genotypes of the offspring. The IDDM-associated allele *PstI* a (1,127/*PstI*) was transmitted to 38 of 50 HLA-DR4-positive diabetic offspring in informative male meioses ($P < 0.0005$), whereas 8 of 21 non-DR4 diabetics received this allele (Table 3b). The effect of maternally transmitted alleles is not significant for any genotype. Data from other loci, including the 5' VNTR with alleles grouped into size classes, showed similar segregation patterns (Table 3b). We thus decided to reanalyse data from

the 5th Genetic Analysis Workshop (GAW) which includes HLA and 5' VNTR genotypes for 94 IDDM families (data obtained from F. Clerget-Darpoux). When we classified the 5' VNTR alleles by their size, the results were remarkably similar to ours (Table 3b).

Previous studies of IDDM susceptibility have focused on insulin as a candidate gene because of its β -cell-specific expression. We have identified all mutations in or near *INS* that could account for a contribution to diabetes susceptibility. As none of the mutations alter the sequence of the protein, an *INS* effect may be due to altered regulation of gene expression. Alternatively, the *IGF2* gene, or an unidentified gene product from this region may be responsible for susceptibility.

The linkage in this study was observed principally in male meioses. Maternal-fetal interaction or genomic imprinting could account for this result. We favour the latter explanation because of the previously documented maternal imprinting of genes in this region (11p15)¹⁷⁻²⁰. *Igf2* is known to be imprinted in the mouse¹⁸ and the *INS-IGF2* synteny is conserved between species¹⁸. In man, loss of maternal imprinting of 11p15 may be implicated in the Beckwith-Wiedemann syndrome^{19,20}, which is associated with both islet cell hyperplasia and hyperinsulinaemia^{21,22}. Maternal imprinting could also account for the increased risk of disease in offspring of diabetic fathers compared with diabetic mothers²³.

Fi
para
dis
the
sus
tran

Receiv
1. Bi
2. Br
3. Hi
4. Fr
5. Cr
6. Tr
7. Cr
8. Si
9. Br
10. Si
11. U
12. B
13. U
14. B
15. H

De
2
ps
Ste
Eliz
Depa
Medi

THE
5α-1
both
and
Sing
map
traci
isola
pros
gene
5α-1
5α-1
chai
of ti
cons
dele
pseu
resu
and
in m

T
fract
hum
exp
scre
was
an e
an i
cou
redu
info
libra
of 5

* To w
NATI



Mailing Certificate

The following items were sent by first class mail with sufficient postage today, Feb. 5, 2007, by me, Robert McGinnis, to Mail Stop RCE, Commissioner for Patents P.O. Box 1450 Alexandria, VA 22313-1450. The items are for application 10/037, 718, art unit 1637, Examiner Horlick, K.

- 1) Amendment/Response and RCE (26 pages) signed
- 2) Enclosures: Selected pages from the McMahon, the Inventor's paper AHG98, Julier, Spielman, and Thomson references, some marked. Total of 15 pages.
- 3) Credit card form PTO-2038 with fee for Small Entity RCE, 1 month extension, extra claims 1 independent, 9 dependent (1 page) signed
- 4) This mailing certificate (1 page) signed

Total pages 43 pages or sheets.

- 5) Return receipt post card (signed)

R. McGinnis

Robert McGinnis
Reg. No. 44, 232
Feb. 5, 2007

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- BLACK BORDERS**
- IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ~~FADING TEXT OR DRAWING~~**
- BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- SKEWED/SLANTED IMAGES**
- COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- GRAY SCALE DOCUMENTS**
- LINES OR MARKS ON ORIGINAL DOCUMENT**
- REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- OTHER: _____**

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.